



Capstone Project 2

# PREDICTING HOUSE PRICE

# Who is the audience?

- ▶ Banks and Financial Investors
- ▶ Real estate company and marketplace



This Photo by Unknown Author is licensed under [CC BY-NC-ND](#)



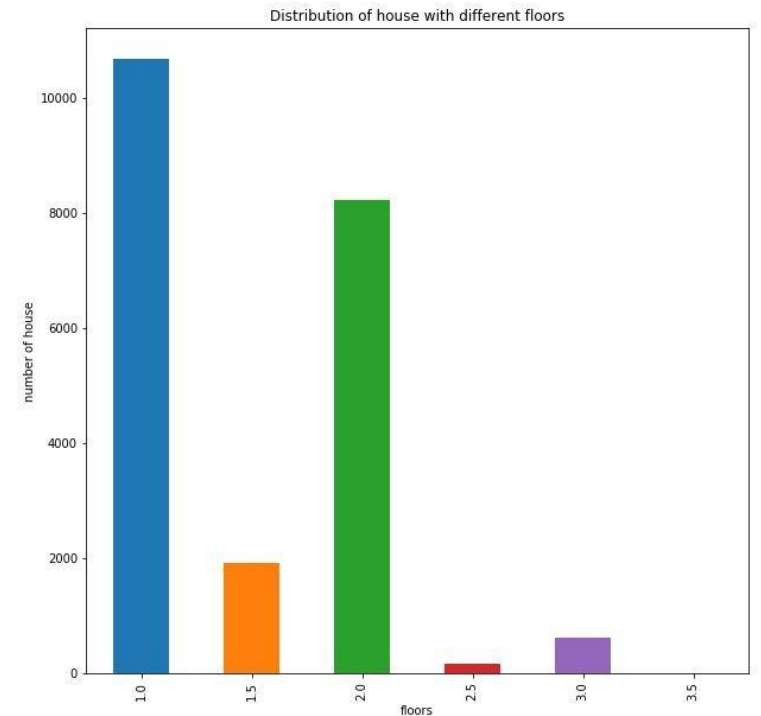
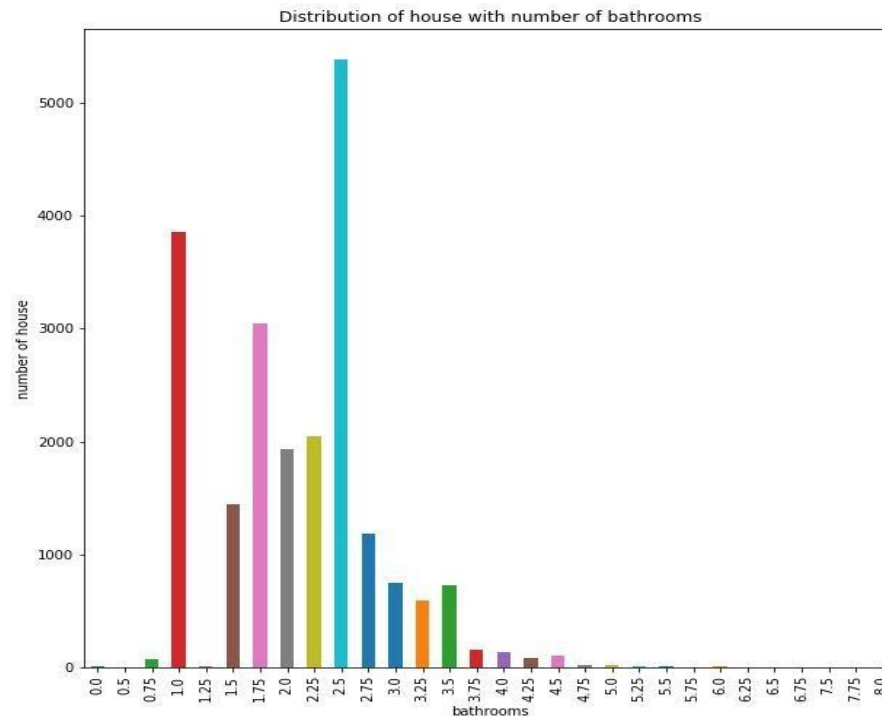
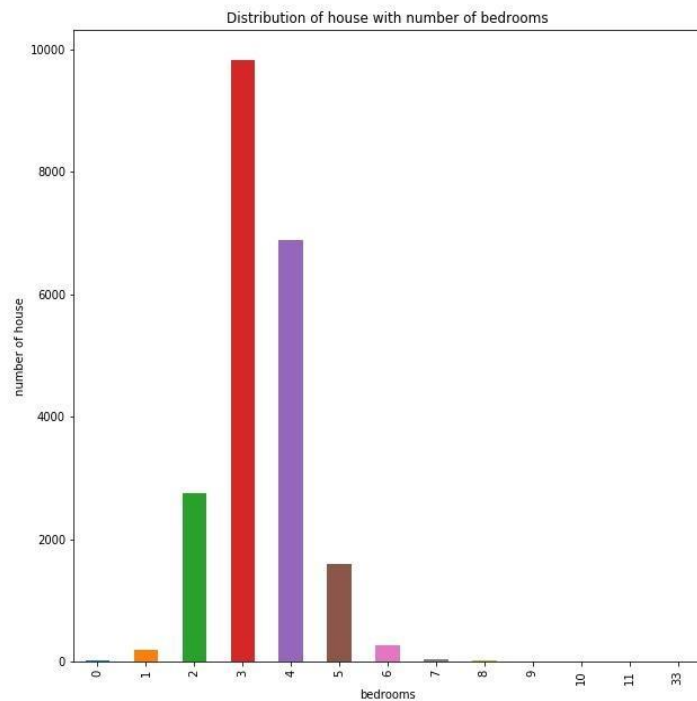
This Photo by Unknown Author is licensed under [CC BY-SA](#)

# Data

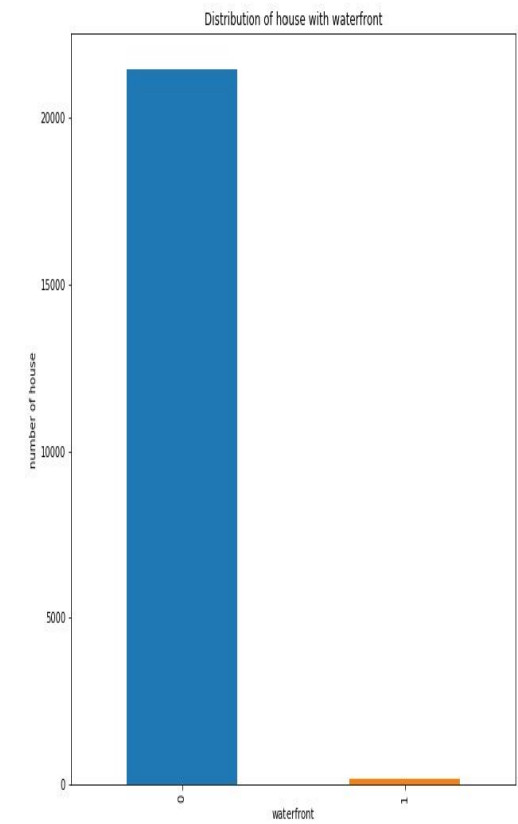
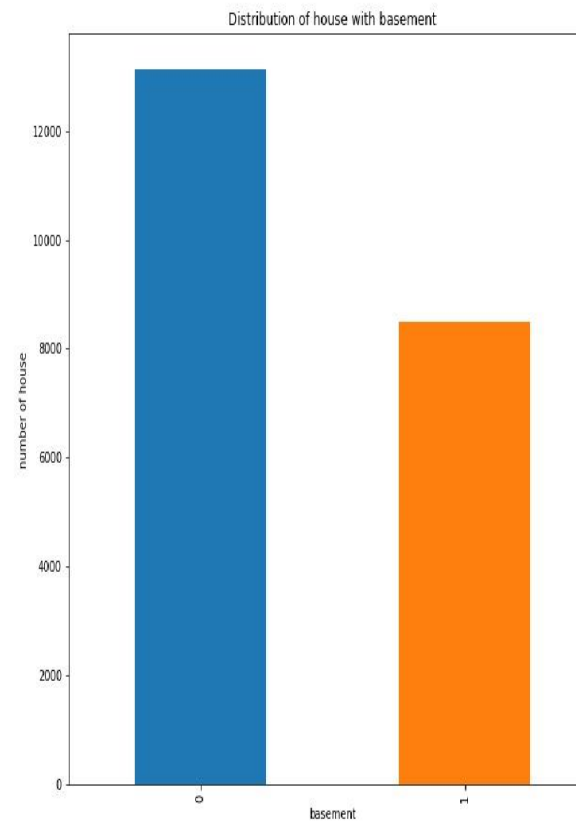
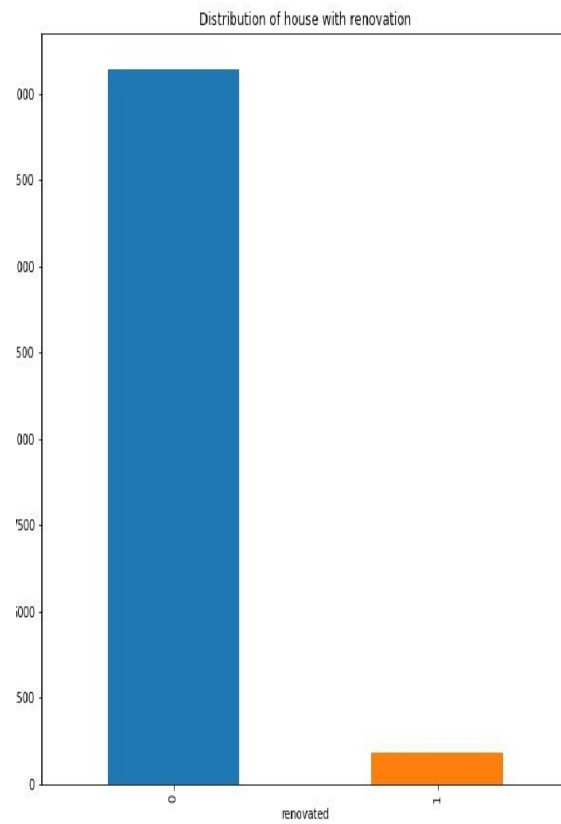
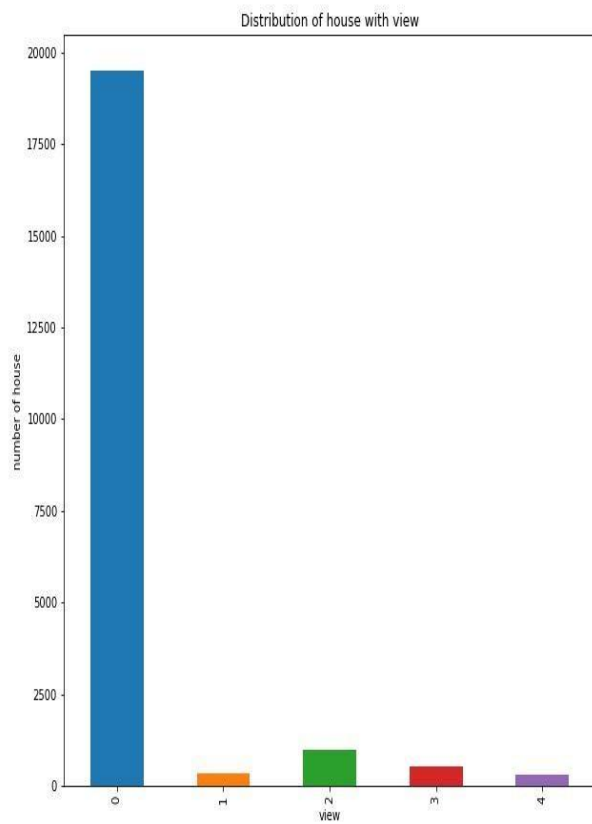
- ▶ Kings County, Seattle, Washington
- ▶ House sold between May 2014 and May 2015
- ▶ 21613 observations and 19 features
- ▶ No missing data
- ▶ There were some outliers

# Exploratory Data Analysis

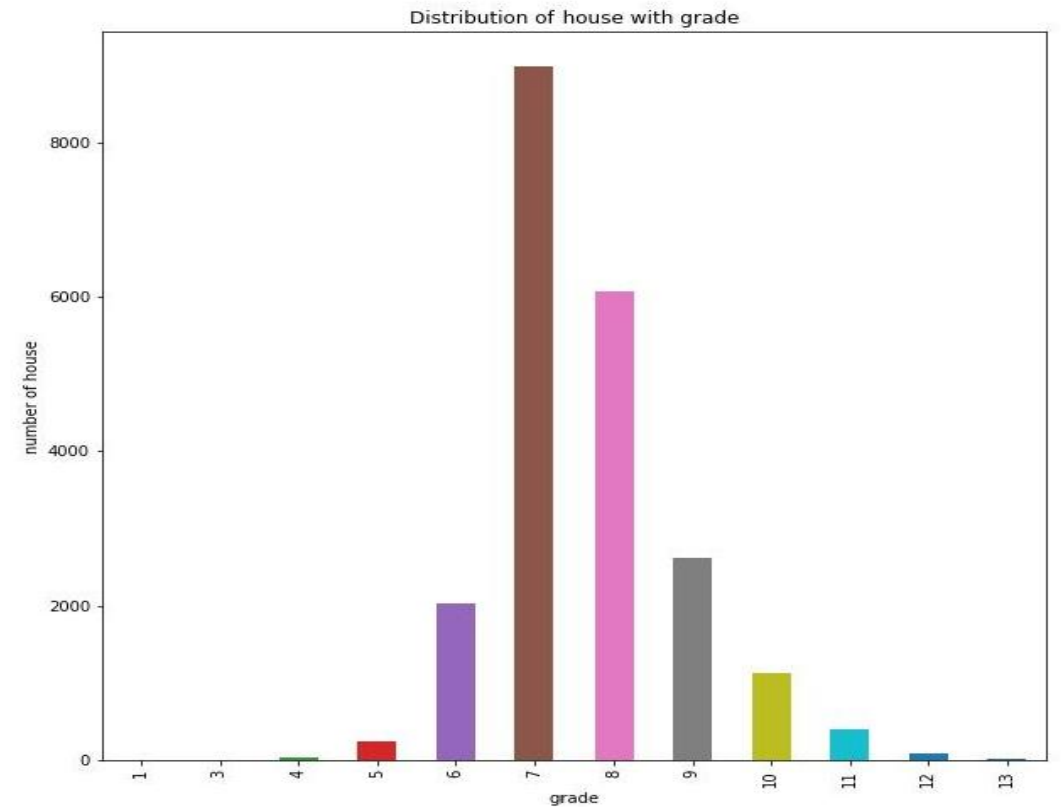
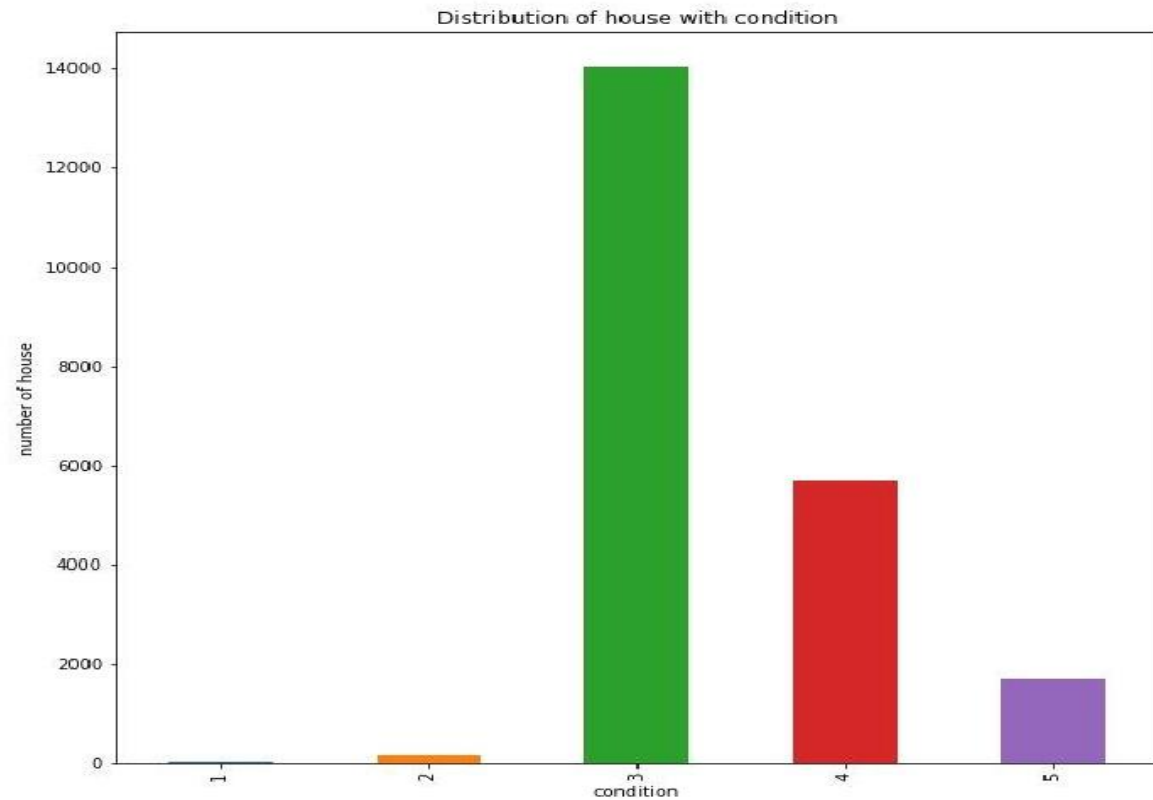
These bar plot will show us what kind of house were most sold in Kings county, Washington.



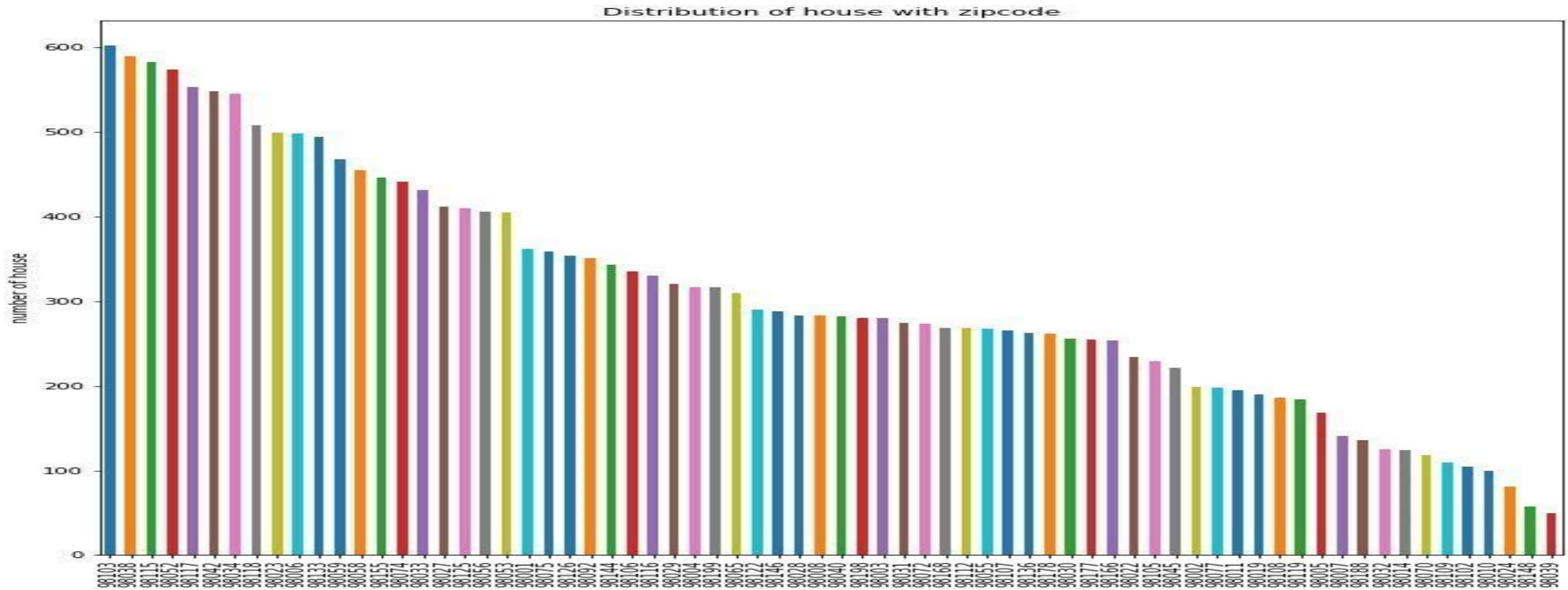
# EDA continued...



# EDA continued..

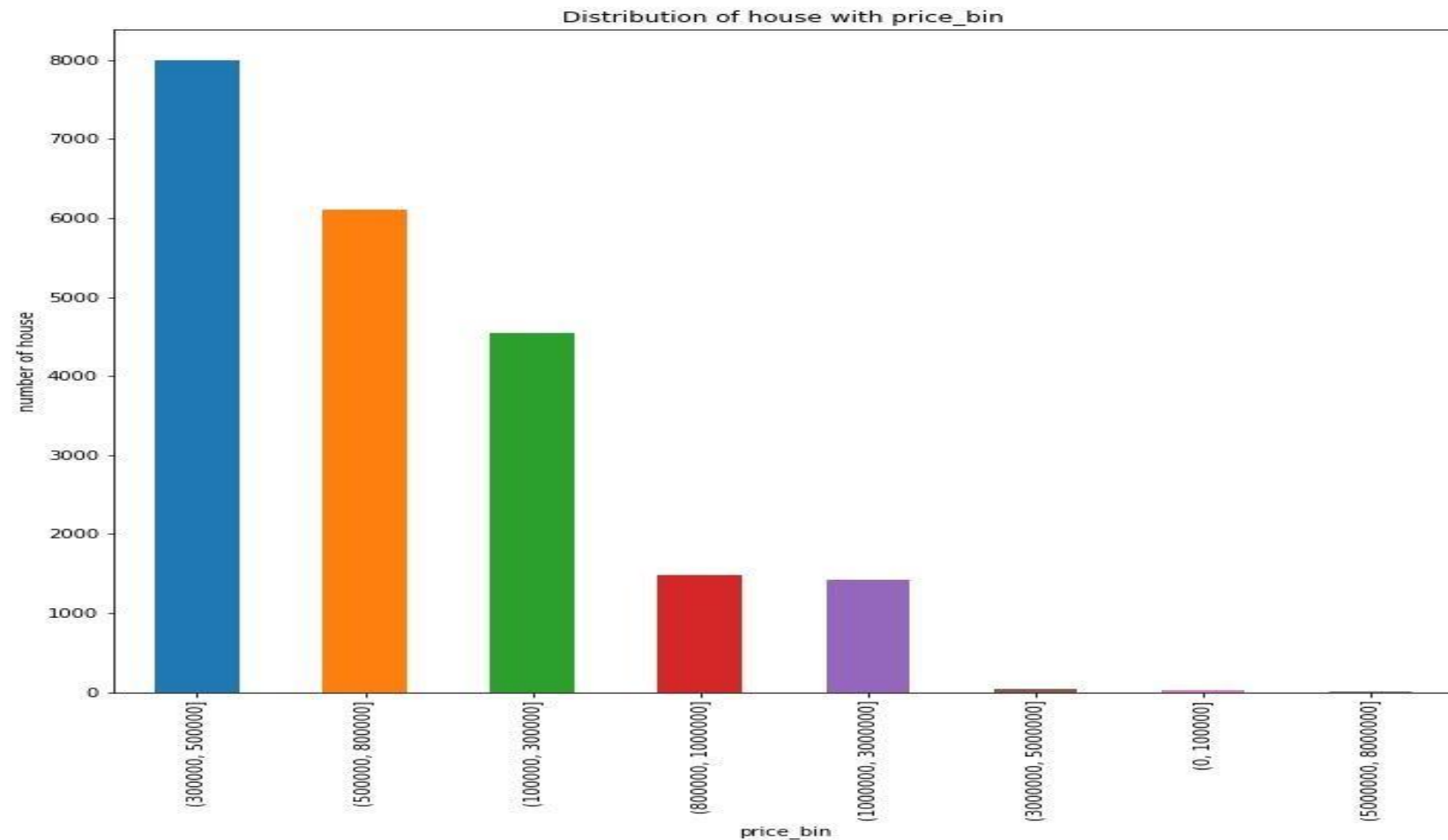


# EDA continued...



# EDA continued...

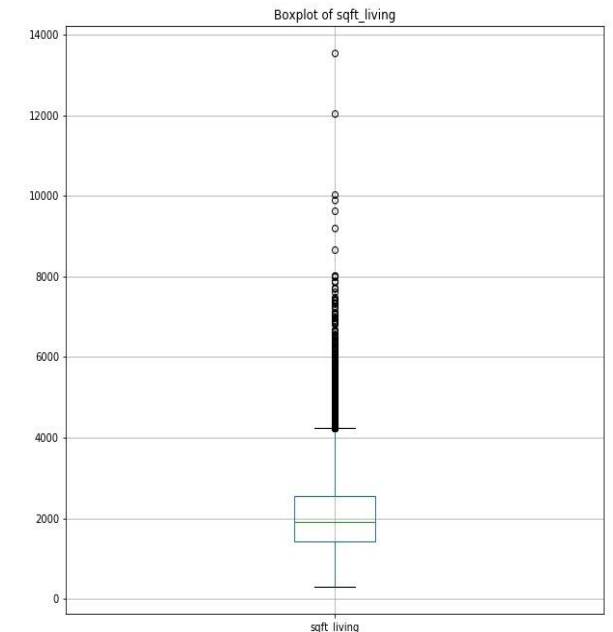
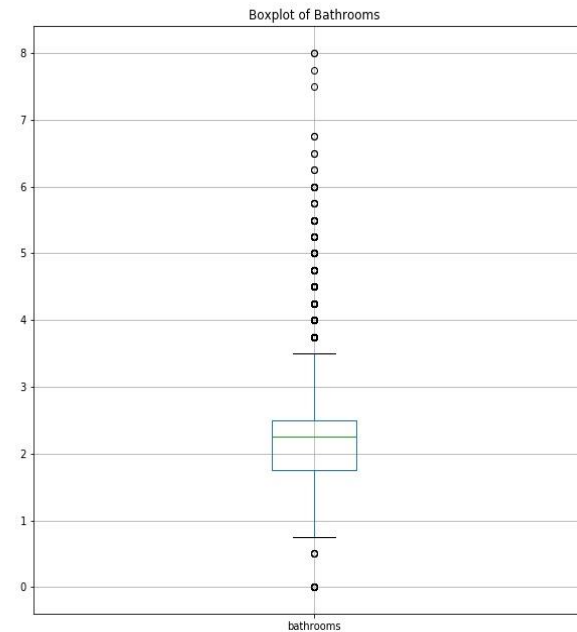
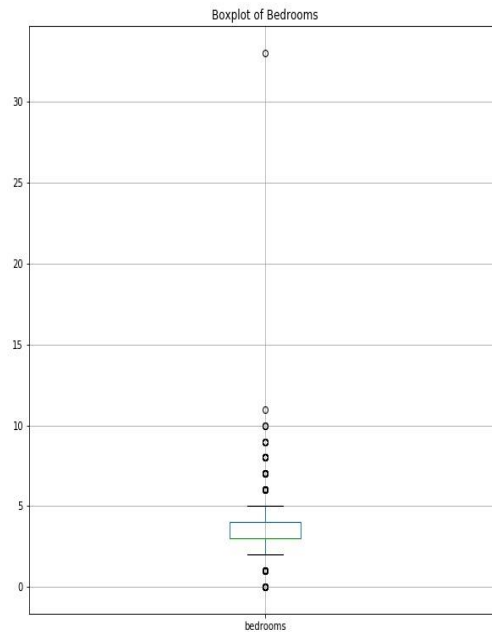
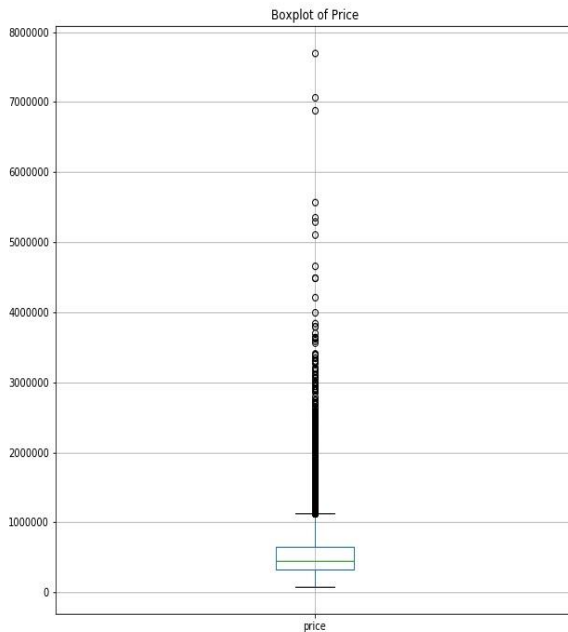
Most of the houses sold are in price-range of 300,000 to 500,000 followed by 800,000.





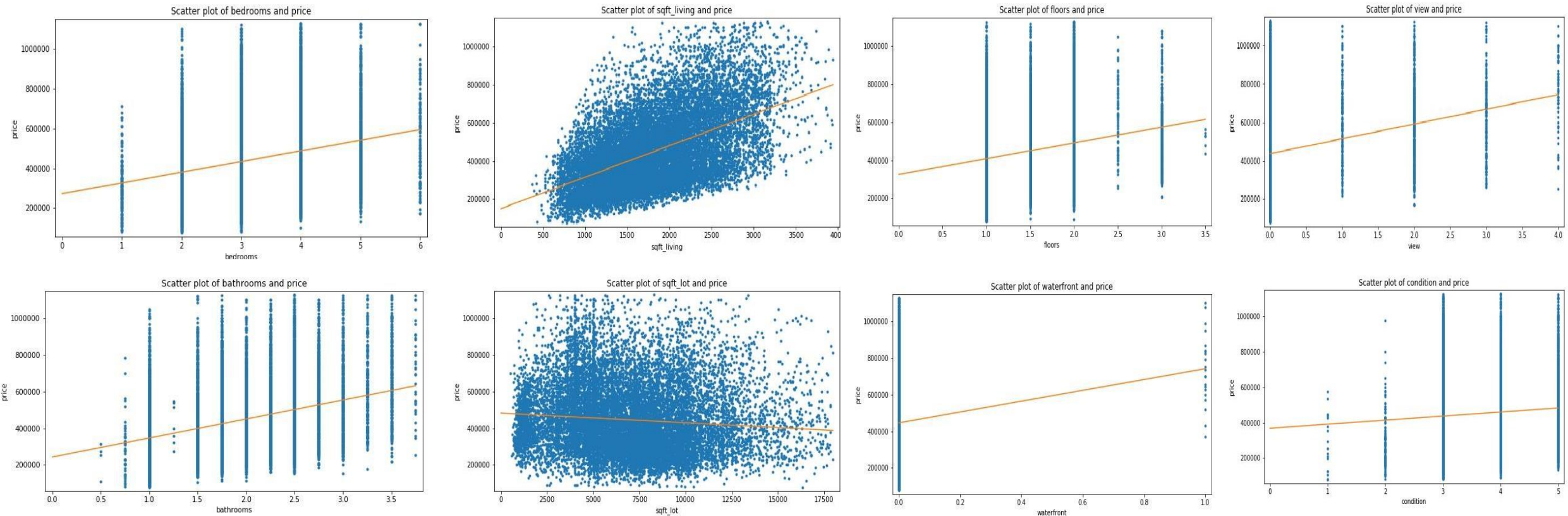
# Data Wrangling or Cleaning.

- ▶ Data must be cleaned and prepared for Machine Learning Model.
- ▶ There were outliers in the dataset, which was dropped or removed.



# Finding Correlation with Price.

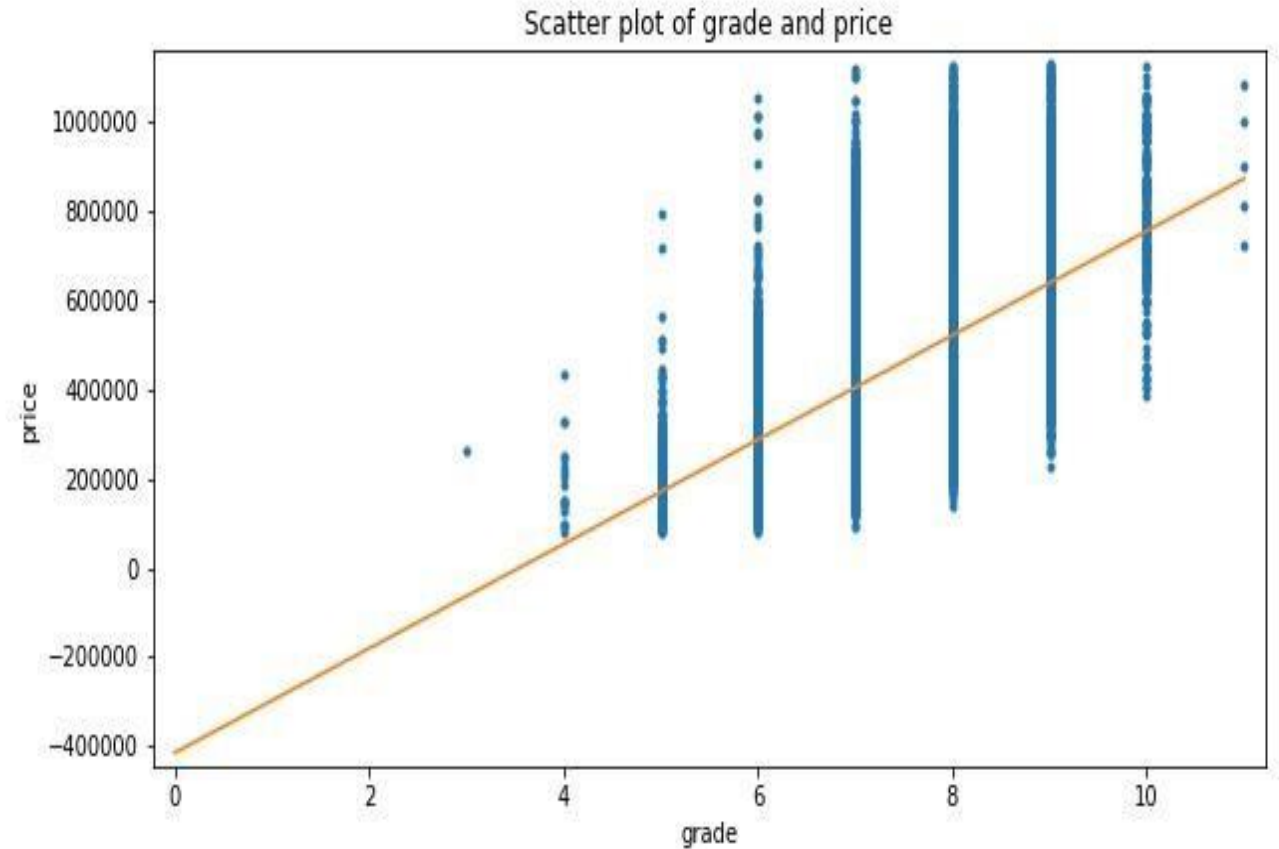
- I used scatter plot to check the correlation of different variables against the price of the house.



# Finding correlation with price.

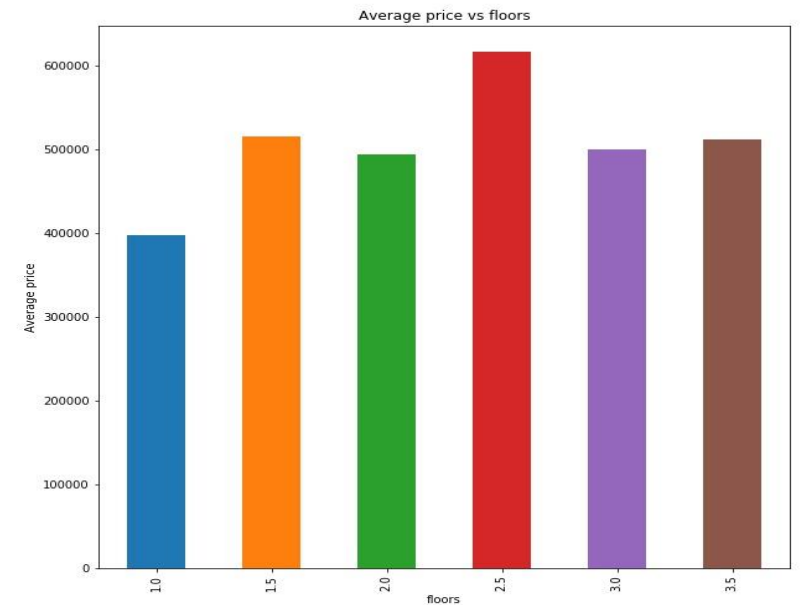
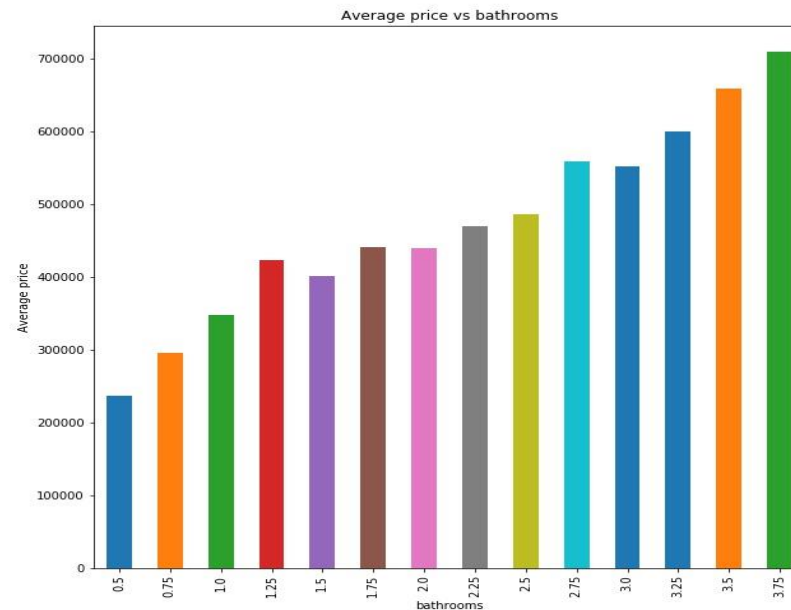
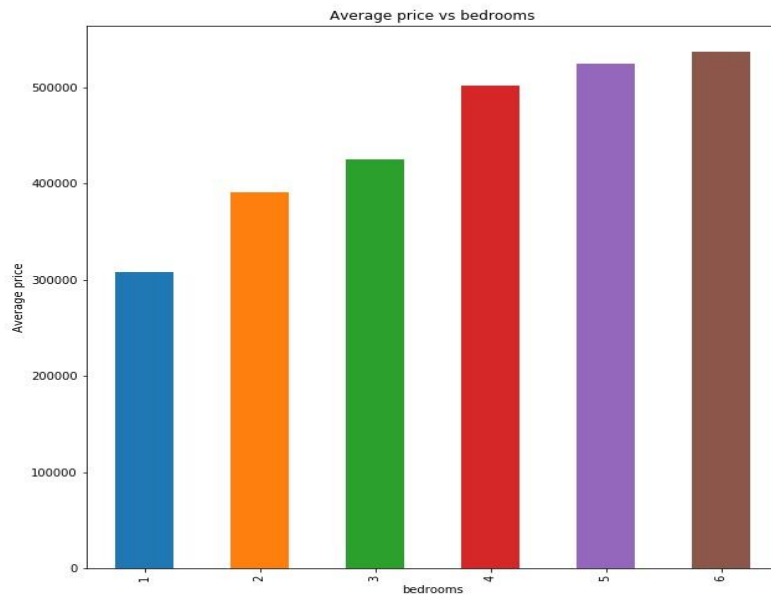
I calculated the correlation coefficient for all the variables to find the best predictors of the house price.

Features	CORRCOEFF	
sqft_lot15	-0.107535	Weak negative
sqft_lot	-0.089069	Weak negative
waterfront	0.055702	Very Weak positive
condition	0.078840	Very Weak positive
view	0.218874	Weak positive
bedrooms	0.235083	Weak positive
floors	0.238493	Weak positive
sqft_basement	0.239227	Weak positive
bathrooms	0.360725	Strong positive
sqft_above	0.403418	Strong positive
sqft_living15	0.439548	Strong positive
sqft_living	0.524052	Strong positive
grade	0.546210	Strong positive

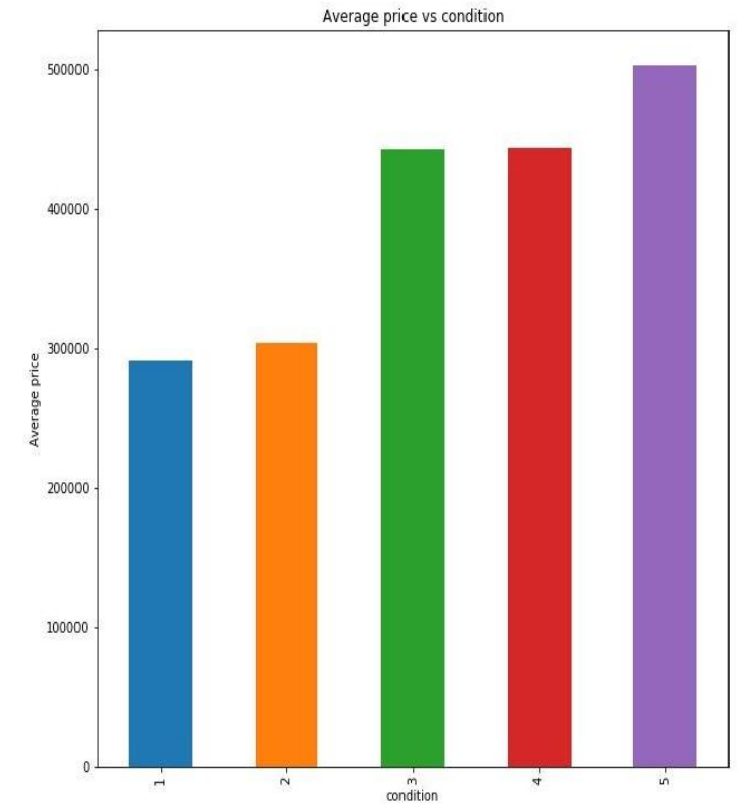
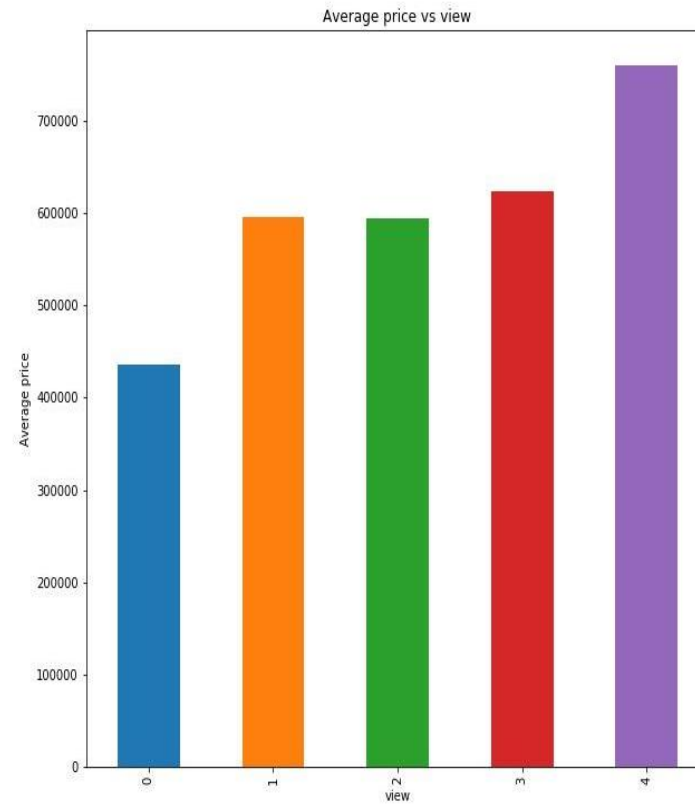
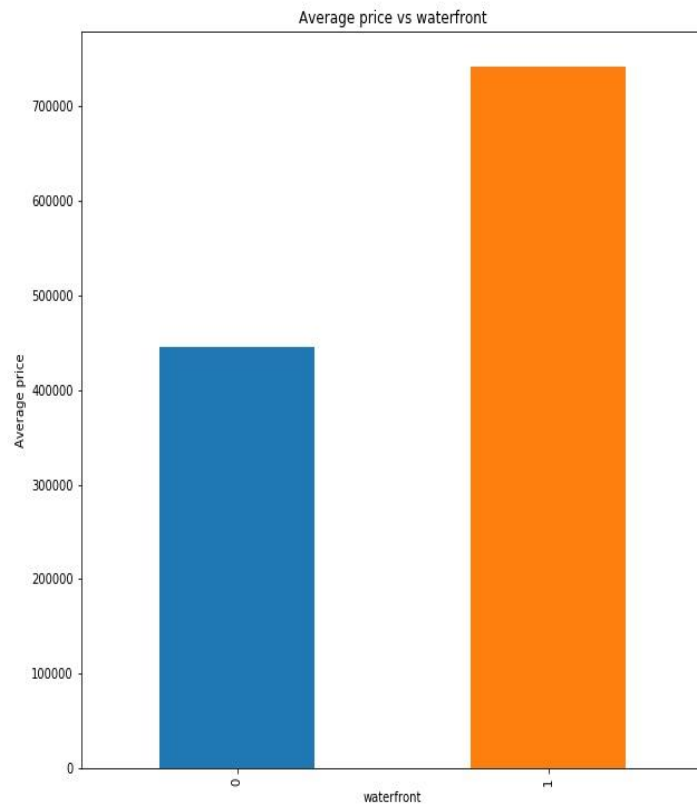


# Data Story

These bar plot can shows us how the average price of the house is affected by some predictor variables.

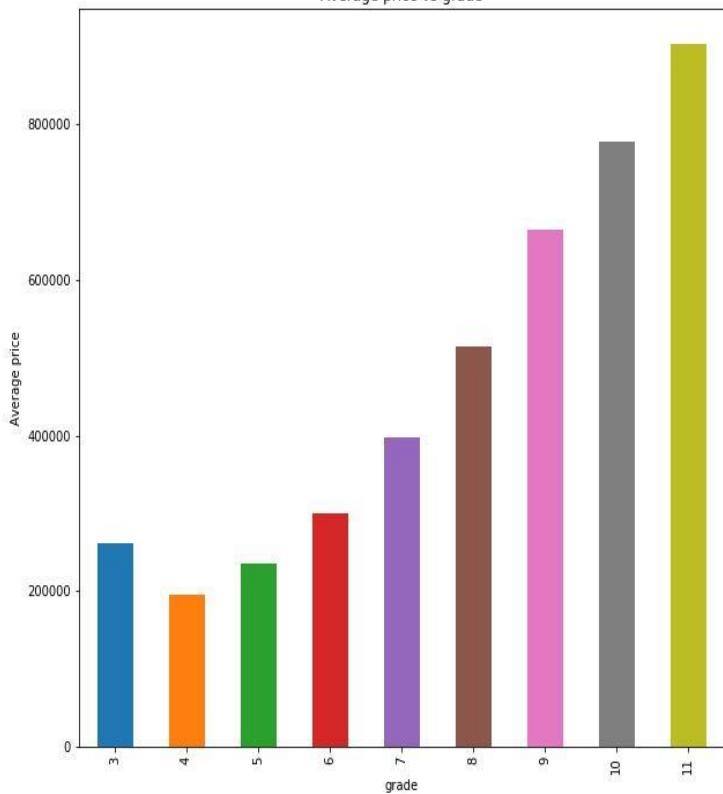


# Data Story Cont...

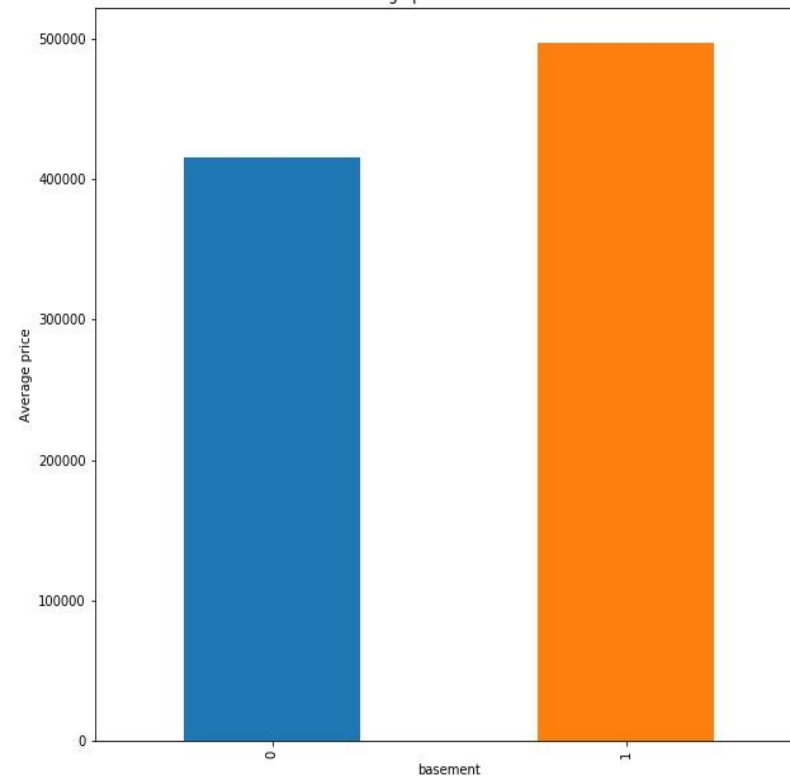


# Data Story...

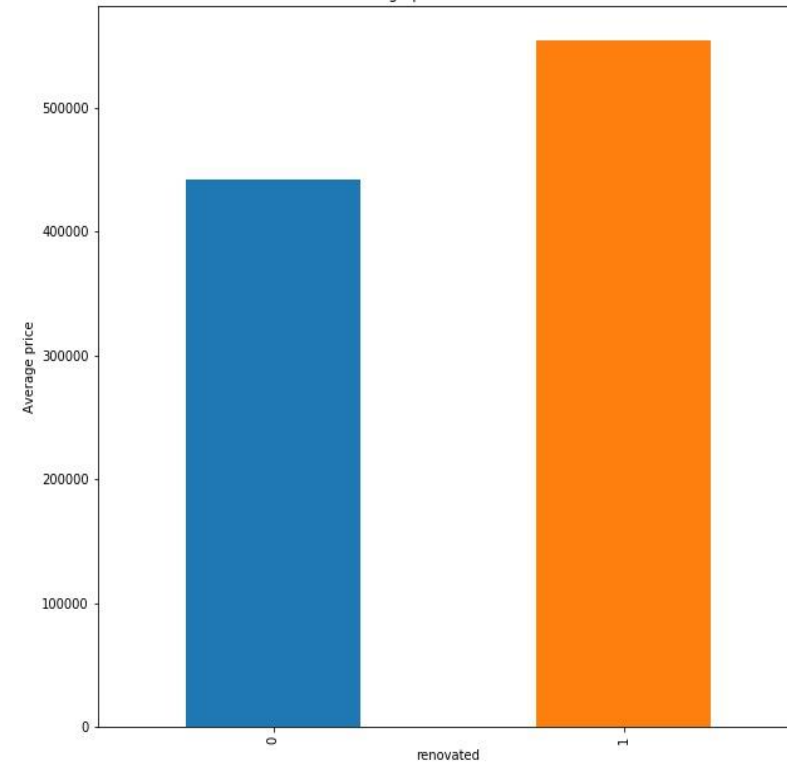
Average price vs grade



Average price vs basement



Average price vs renovated



# Inferential Statistics

- ▶ #H0: There is no significant correlation between number of bedroom and price.
- ▶ #Ha: There is a correlation between number of bedrooms and price.
- ▶ The p-value is less than level of significance 0.05, so we reject the null hypothesis. There is a correlation between number of bedrooms and price.
- ▶ I performed the hypothesis testing to check if the correlation between price and other features happened by chance.

# Machine Learning

- ▶ Linear Regression
- ▶ Decision Tree Regressor
- ▶ Gradient Boosting Regressor
- ▶ Random Forest Regressor

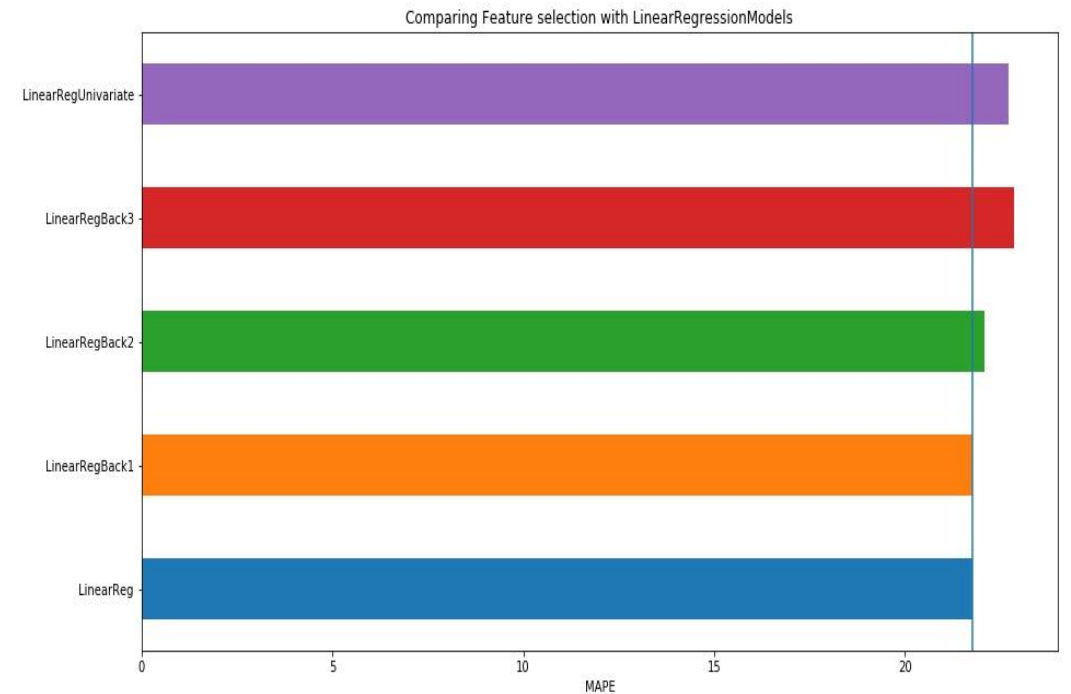
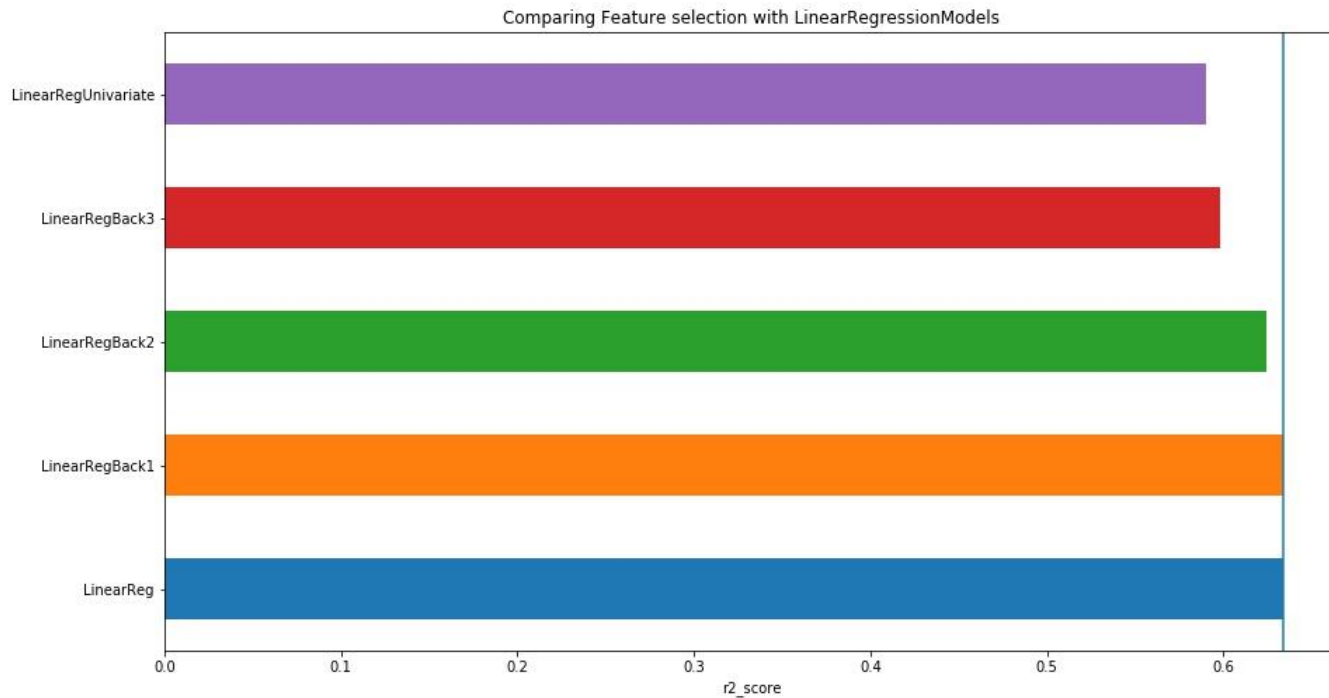


# Metrics

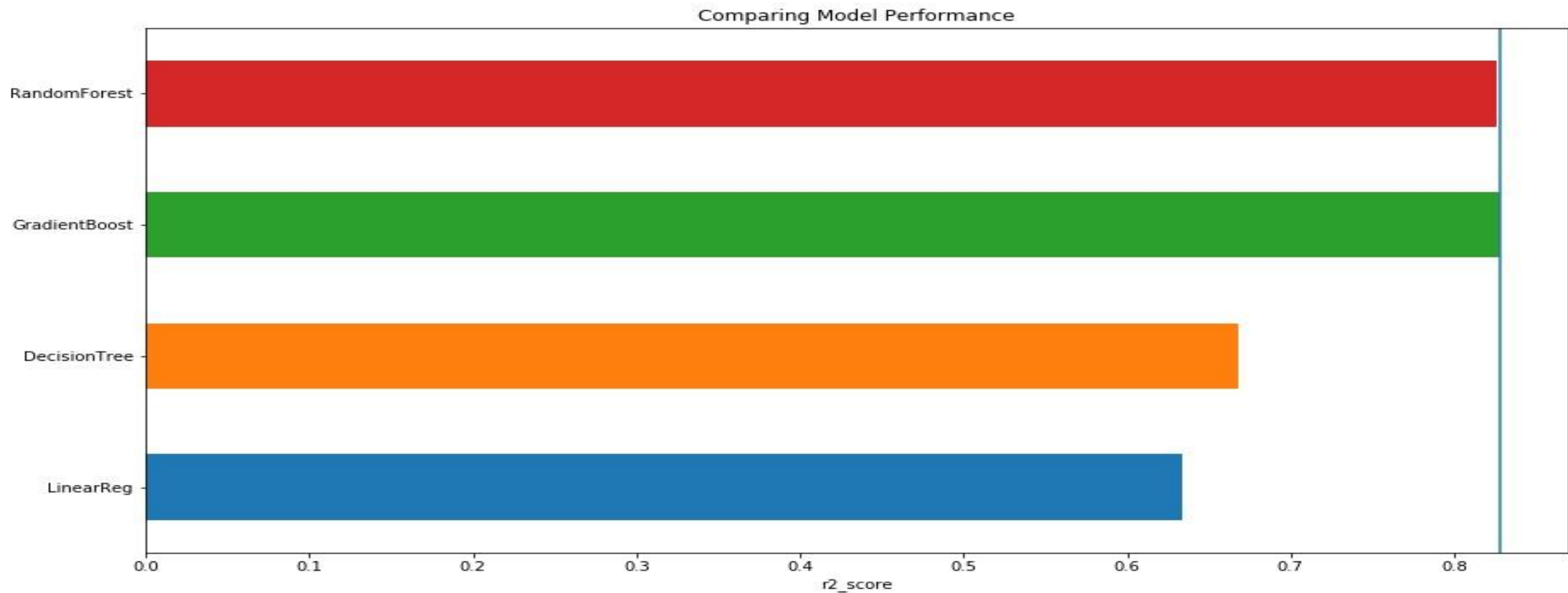
- ▶ Mean Squared Error (MSE)
- ▶ Root Mean Squared Error (RMSE)
- ▶ R2\_Score
- ▶ Mean Absolute Error (MAE)
- ▶ Mean Absolute Percent Error (MAPE)

# Using Feature Selection and Compare Model's performance.

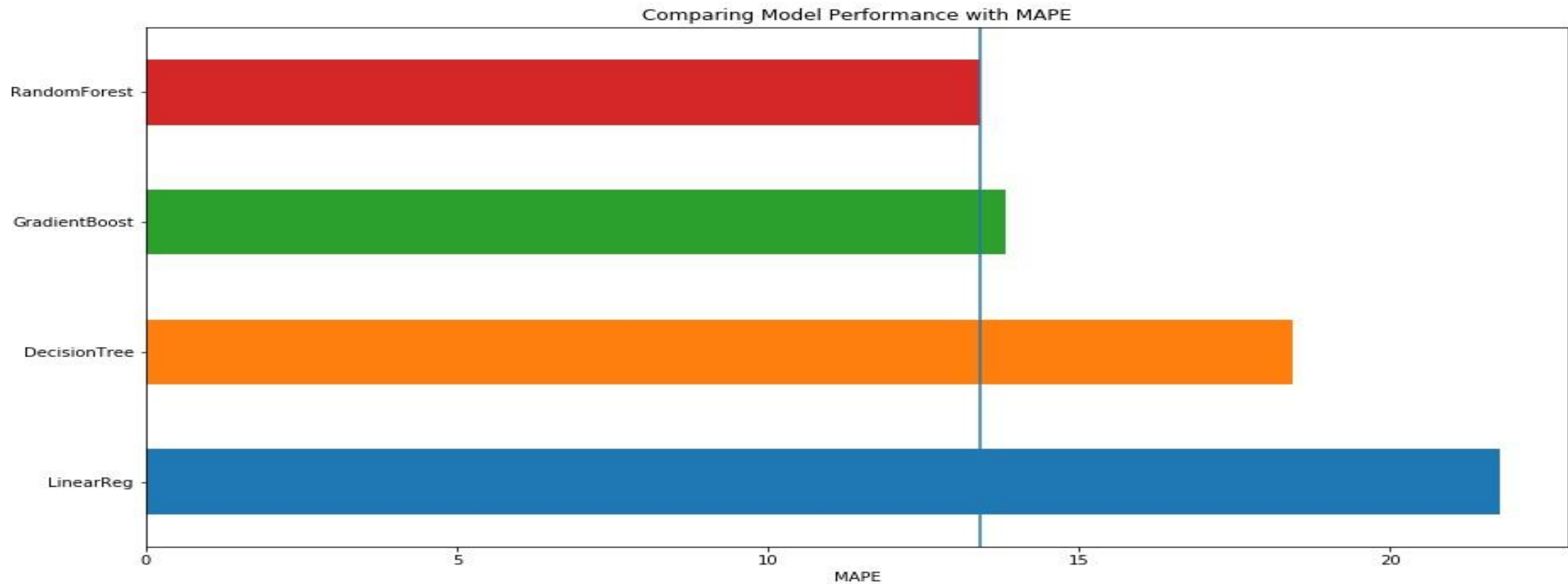
- ▶ Backward Elimination
- ▶ Univariate Elimination



# Compare Different Regressor Models

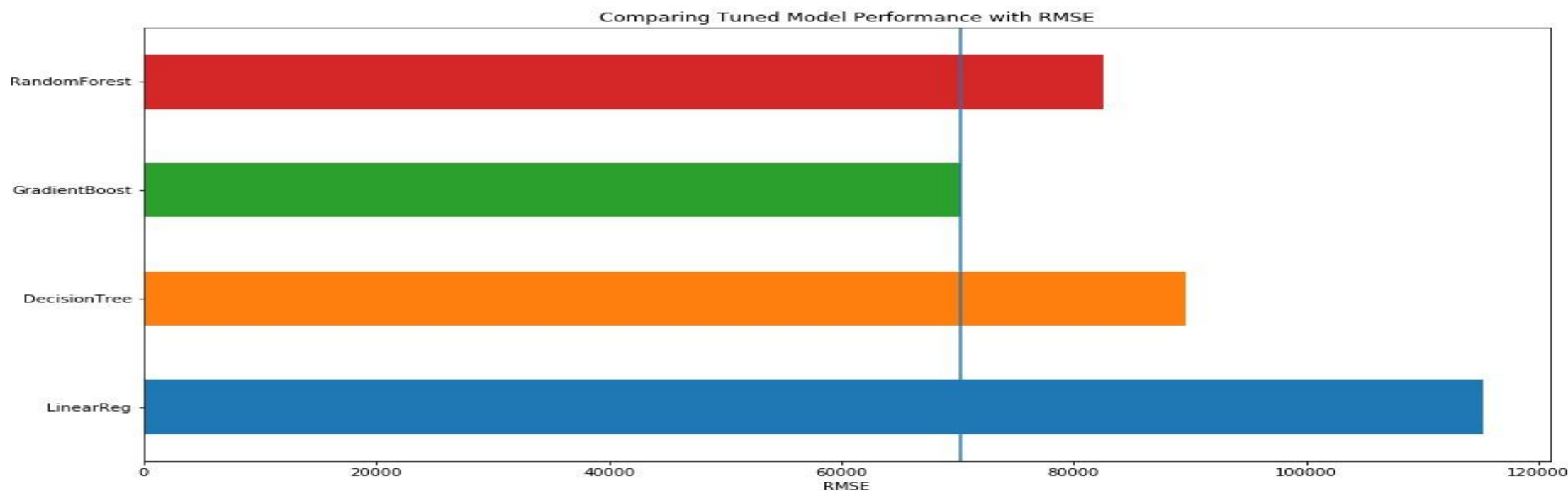


# Comparing different Regressor Models

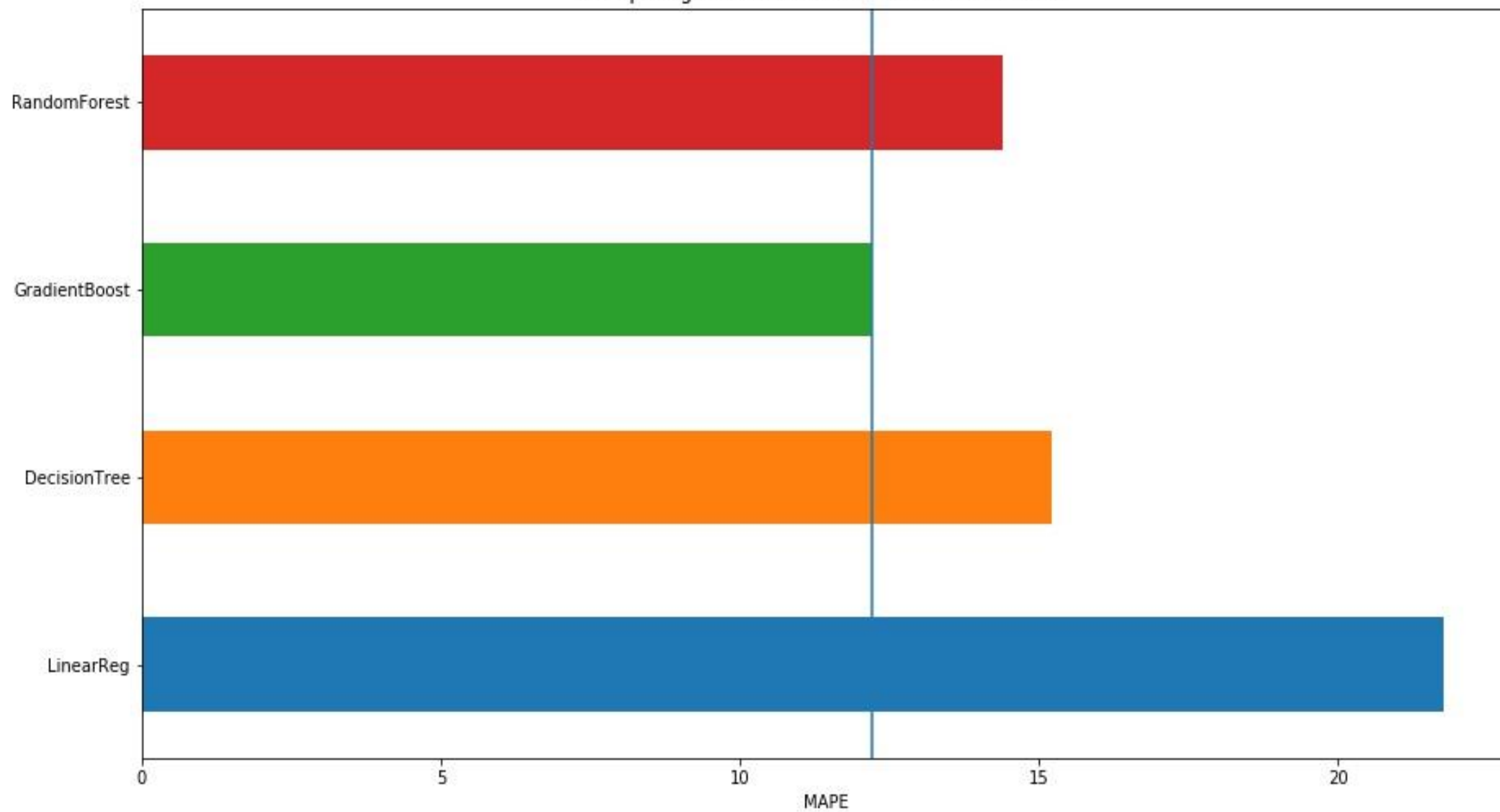


# Hyper-parameter Tuning and Comparing Tuned Model's Performance

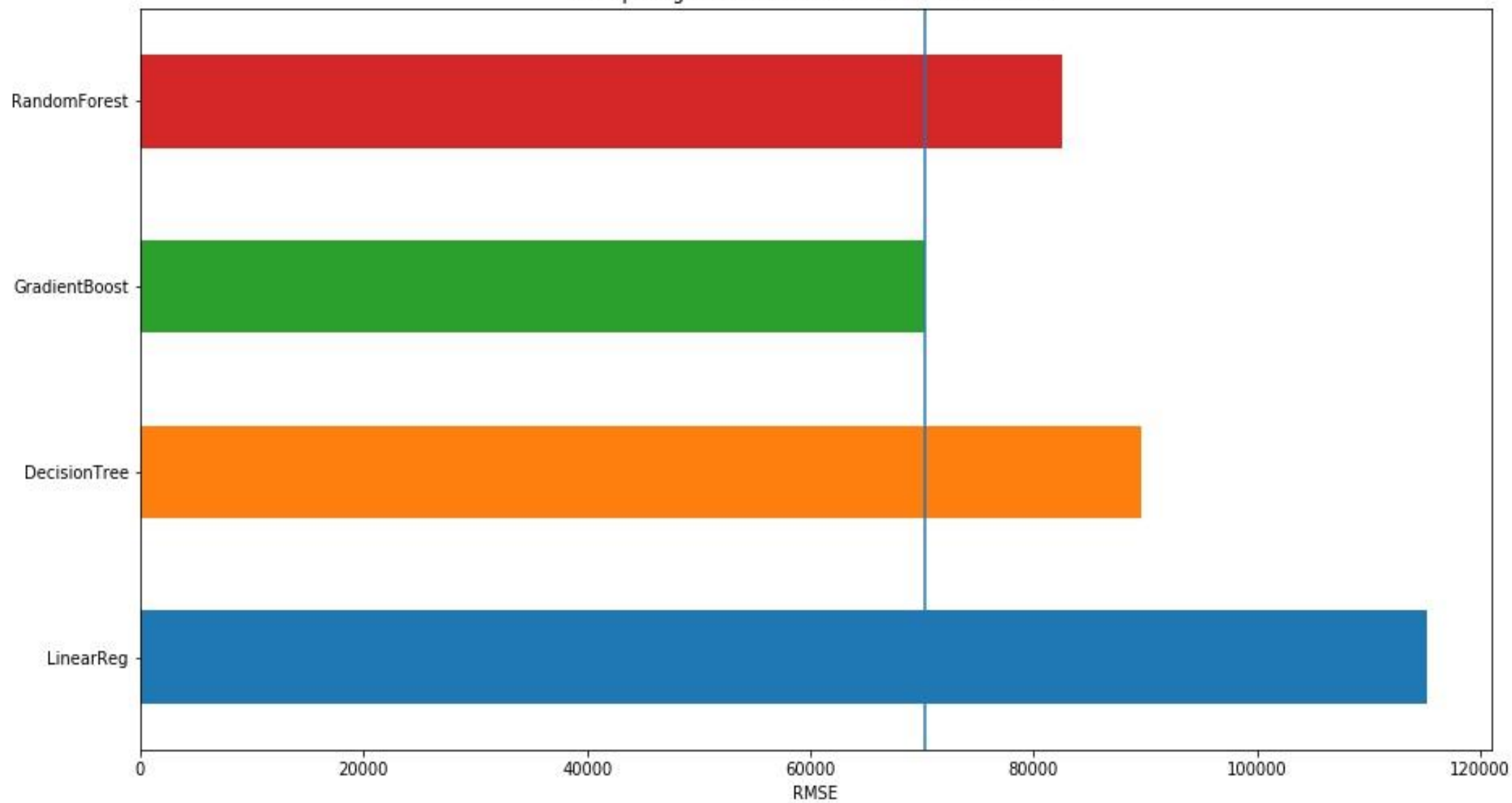
- ▶ GridSearchCV
- ▶ RandomizedSearchCV



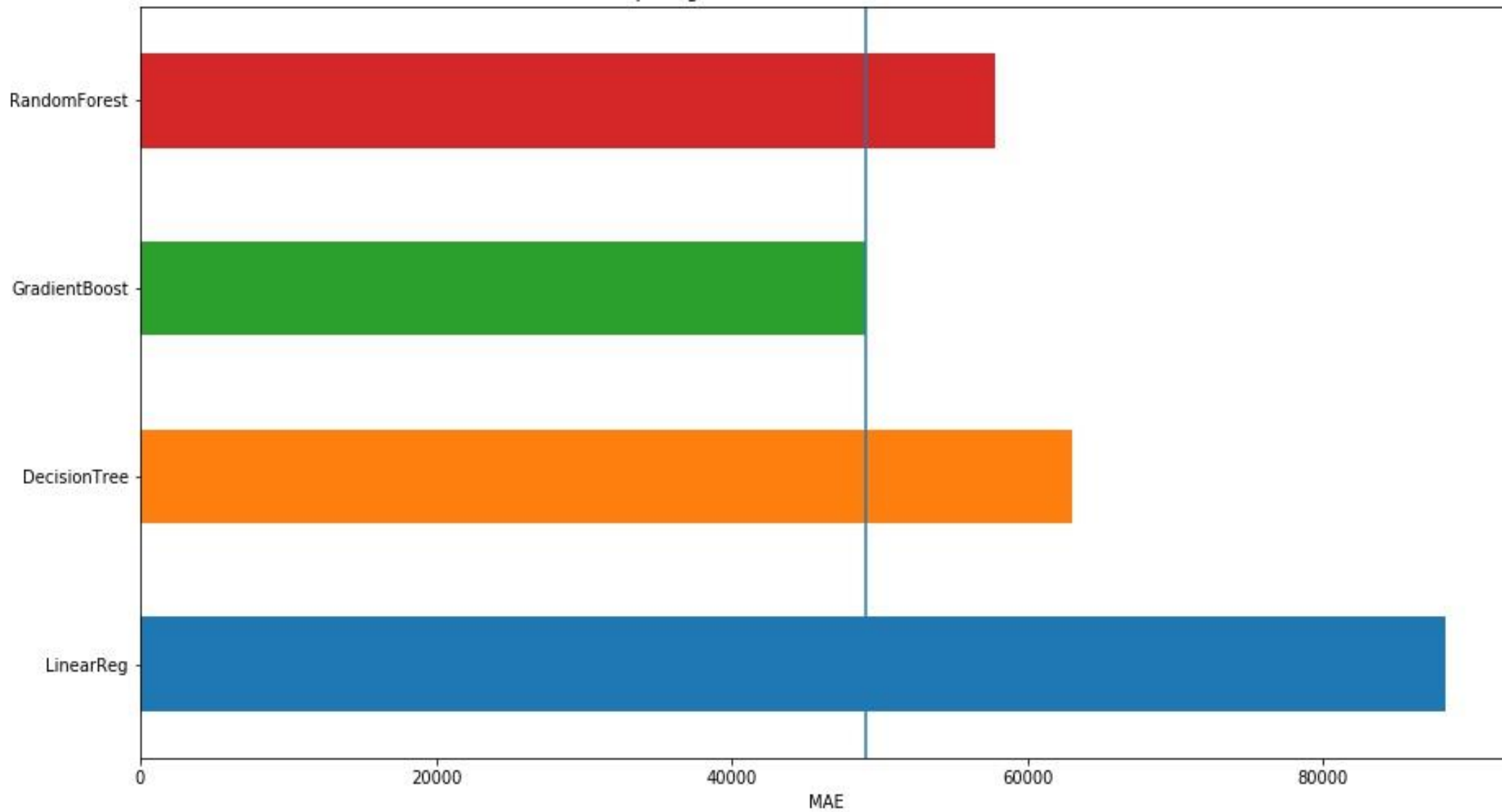
Comparing Tuned Model Performance with MAPE



Comparing Tuned Model Performance with RMSE



Comparing Tuned Model Performance with MAE





# Conclusion

- ▶ The Gradient Boosting Regressor Model is better than random guess, and it is a better performing model compared to other 3 models.
- ▶ In future, we can build other models and compare the performance with this model.
- ▶ Using this model we can predict the house price.
- ▶ This information can be used as a good estimate.